A Web Tool for Building Parallel Corpora of Spoken and Sign Languages

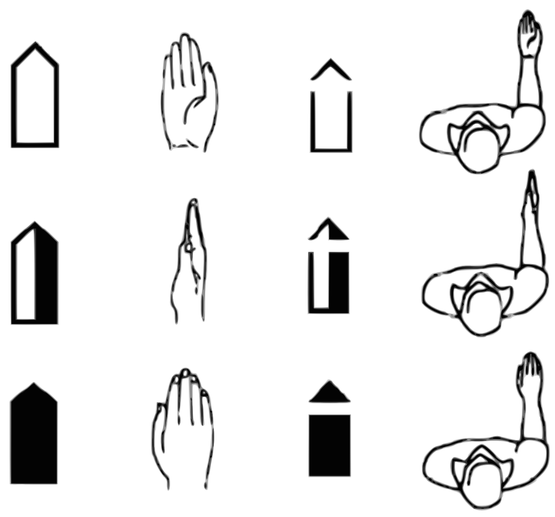
Alex Becker
alex@porthal.com.br
UNIPAMPA

Fabio Kepler
fabio@kepler.pro.br
UNIPAMPA
L^{2}F, INESC-ID

Sara Candeias
t-sacand@microsoft.com
MSFT LDC

Sign Languages

- Over 200 distinct sign languages in the world.
- 70 million deaf people over the world.
- 5.7 million people with hearing impairment in Brazil.
- Children who lose hearing before beginning to speak have a sign language as their native language.
- Among several proposals for writing sign languages, the most prominently is the SignWriting.
- The SignWriting system defines sets of symbols for handshapes, facial expressions, body locations, orientation, contact, and movement.



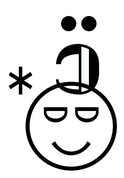
Objectives

- To build an online tool for manual annotation of texts in any spoken language with SignWriting in any sign language.
- To allow the creation of parallel corpora between spoken and sign languages.
- To design it in a way that it eases the task of human annotators by giving smart suggestions as the annotation progresses.
- A parallel corpus between English and American Sign Language could be used for training Machine Learning models for automatic translation between the two languages.

SignWriting Representation

- Signs stored as images have limited applicability.
- Formal SignWriting (FSW) is the latest format for encoding signs.
- FSW encodes logographic words (signs) as strings.

M518x517S16d10494x467S33e00482x482S31b00482x482S21900496x456S20500475x476

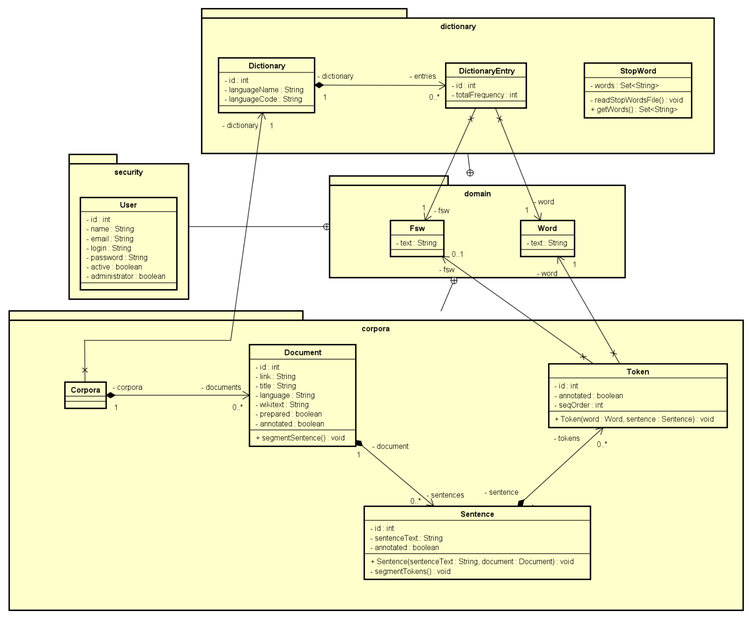


A Spoken-Sign Corpus Annotation Tool

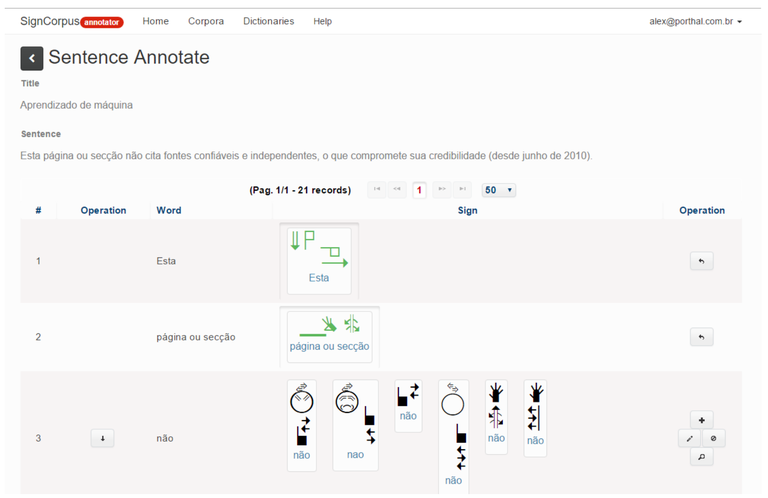
- Uses SignWriting and an existing tool for constructing new signs.
- Supports multiple sign and spoken languages.
- Allows collaborative annotation.
- Provides annotation suggestions based on previous annotations.
- Supports importing an initial dictionary from the SignPuddle portal.

Design and Implementation

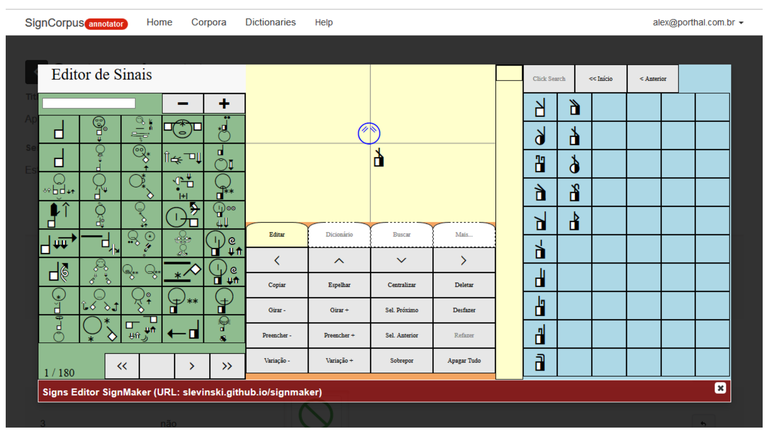
- Java Web platform.
- EJB Application (Enterprise JavaBeans).
- JSF framework (Java Server Faces).
- MVC architecture (Model-View-Controller).



Domain diagram.



Screenshot of the sentence annotation interface.



Screenshot of the SignMaker editor interface.

Final Remarks and Future Work

- Helping the development of proper resources for sign languages that can then be used in state-of-the-art models currently used in tools for spoken languages.
- Open source: https://bitbucket.org/unipampa/signcorpus.
- Next step is to improve the searching and ranking of candidate signs by considering word inflections and by building language models for sign sentences.

This work was partially supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project UID/CEC/50021/2013.